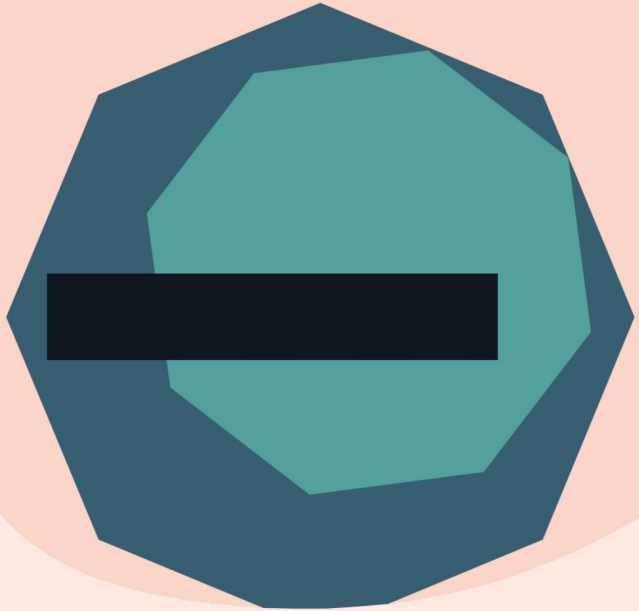
Data Science and the Data Scientist Toolkit

// FLATIRON SCHOOL

Agenda

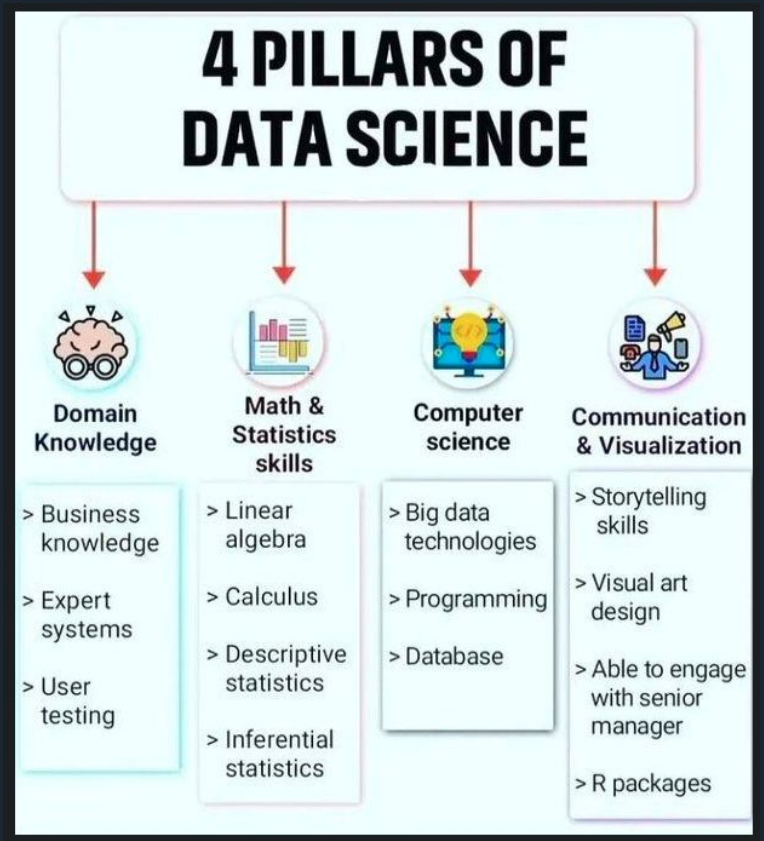
- What is Data Science?
  - Roles and Responsibilities
  - The Process
- The Data Science Toolkit (Phase 1)

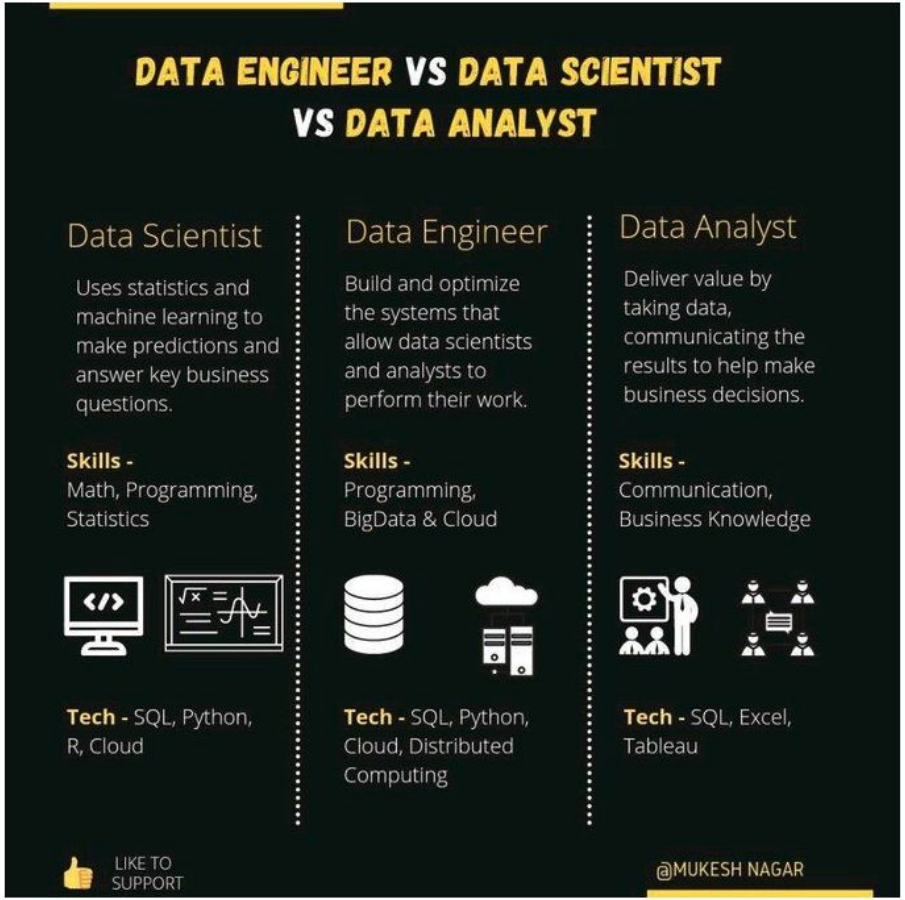
What is Data Science?



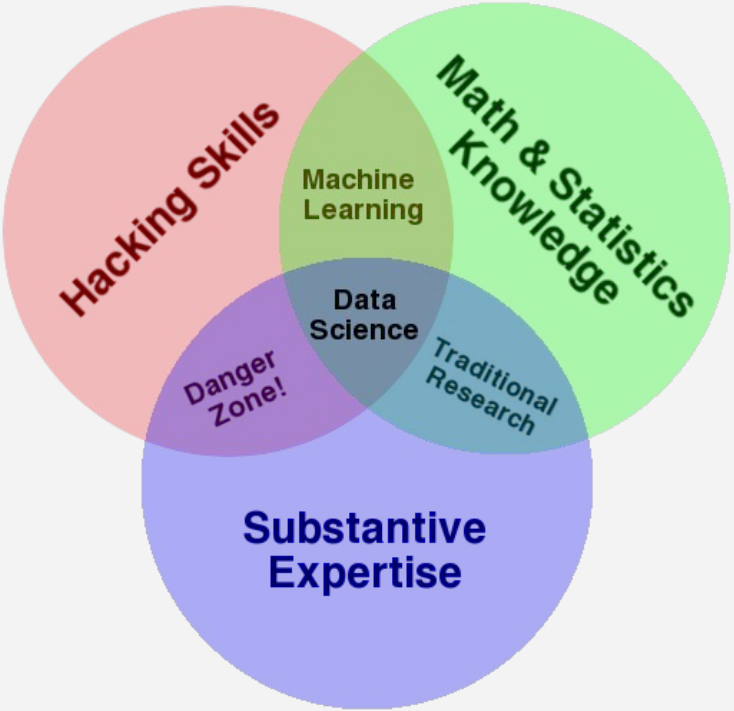
Unique Intersection of Skills

1. Business - Domain Knowledge
2. Math & Statistics - Hypothesis Testing
3. Computer Science - Programming
4. Communication & Presentation - Visualization

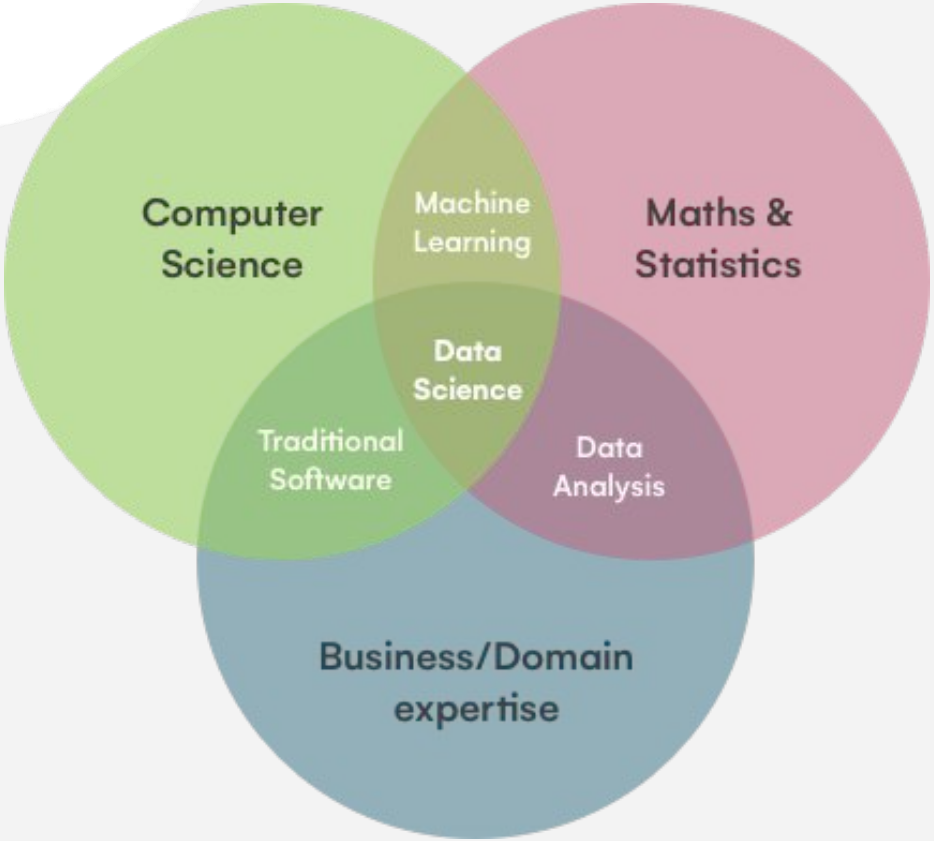



DATA ENGINEER VS DATA SCIENTIST
VS DATA ANALYST
Data Scientist
Uses statistics and machine learning to make predictions and answer key business questions.
Skills -
Math, Programming, Statistics
Tech - SQL, Python, R, Cloud
Data Engineer
Build and optimize the systems that allow data scientists and analysts to perform their work.
Skills -
Programming, BigData & Cloud
Tech - SQL, Python, Cloud, Distributed Computing
Data Analyst
Deliver value by taking data, communicating the results to help make business decisions.
Skills -
Communication, Business Knowledge
Tech - SQL, Excel, Tableau
LIKE TO SUPPORT
@MUKESH NAGAR

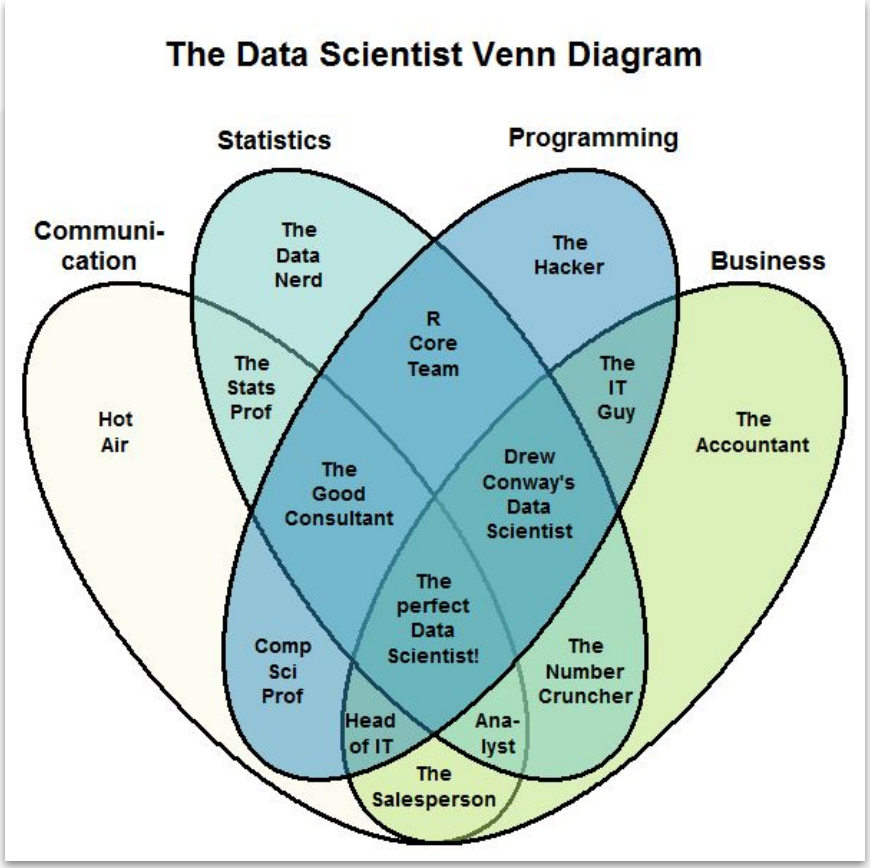
The Data Science Venn Diagram



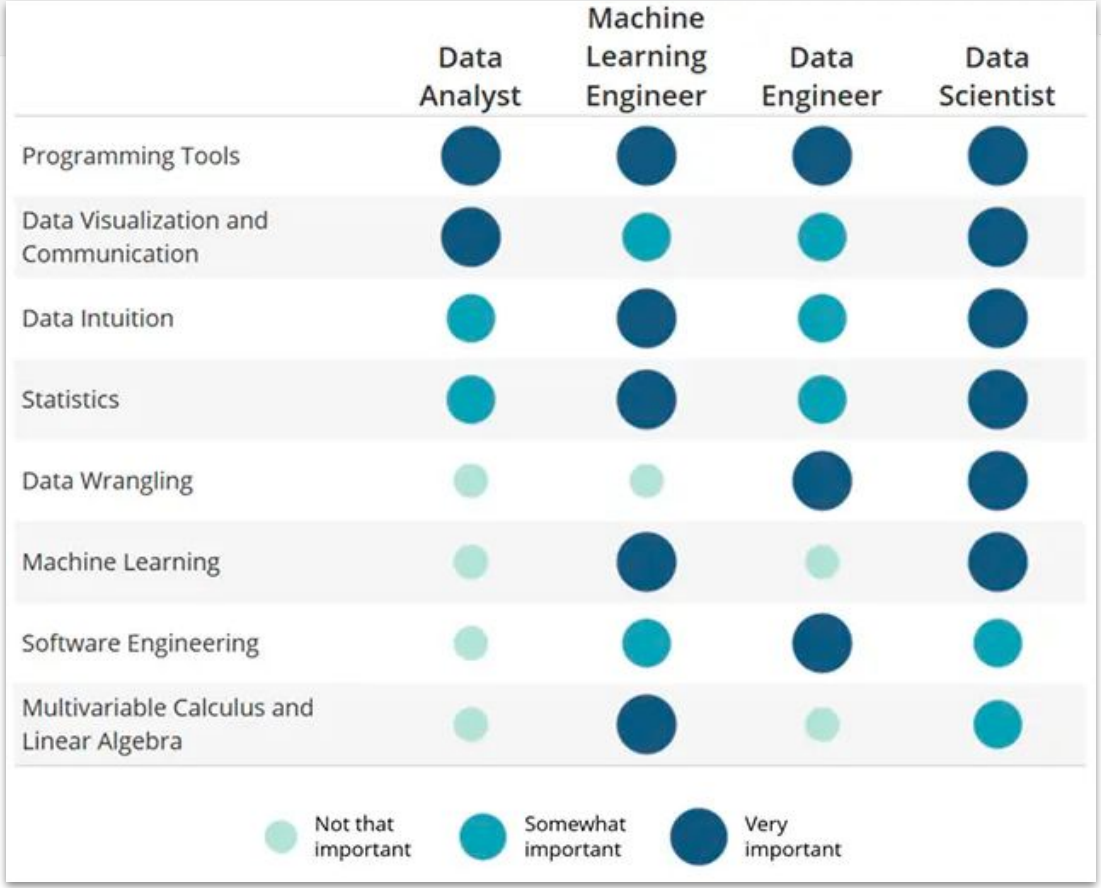
Another Version



And Another

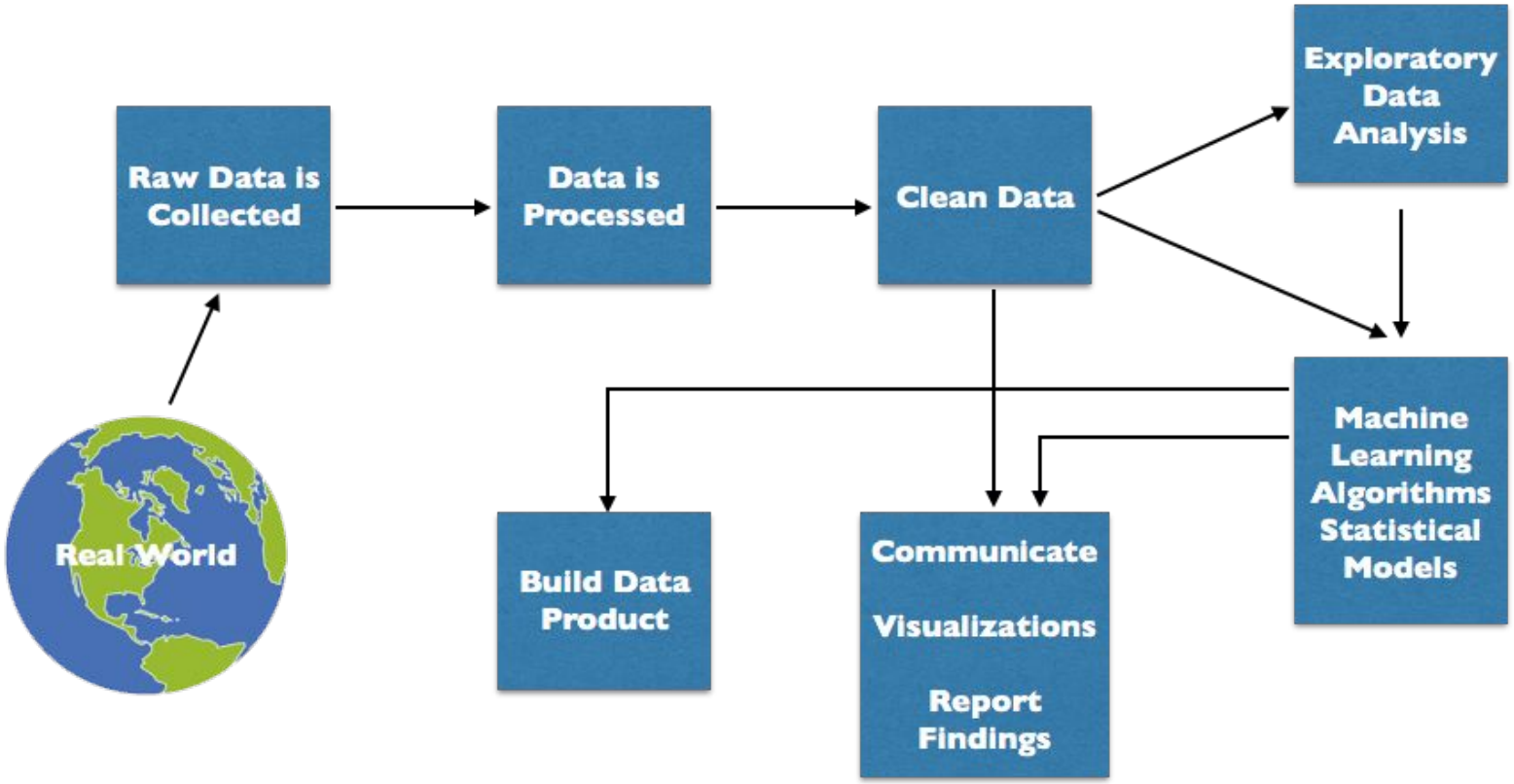

The Data Scientist Venn Diagram
Statistics
Programming
Communi-
cation
Business
The
Data
Nerd
The
Hacker
R
Core
Team
The
Stats
Prof
The
IT
Guy
Hot
Air
The
Accountant
The
Good
Consultant
Drew
Conway's
Data
Scientist
The
perfect
Data
Scientist!
Comp
Sci
Prof
The
Number
Cruncher
Head
of IT
Ana-
lyst
The
Salesperson

Common Roles & Responsibilities

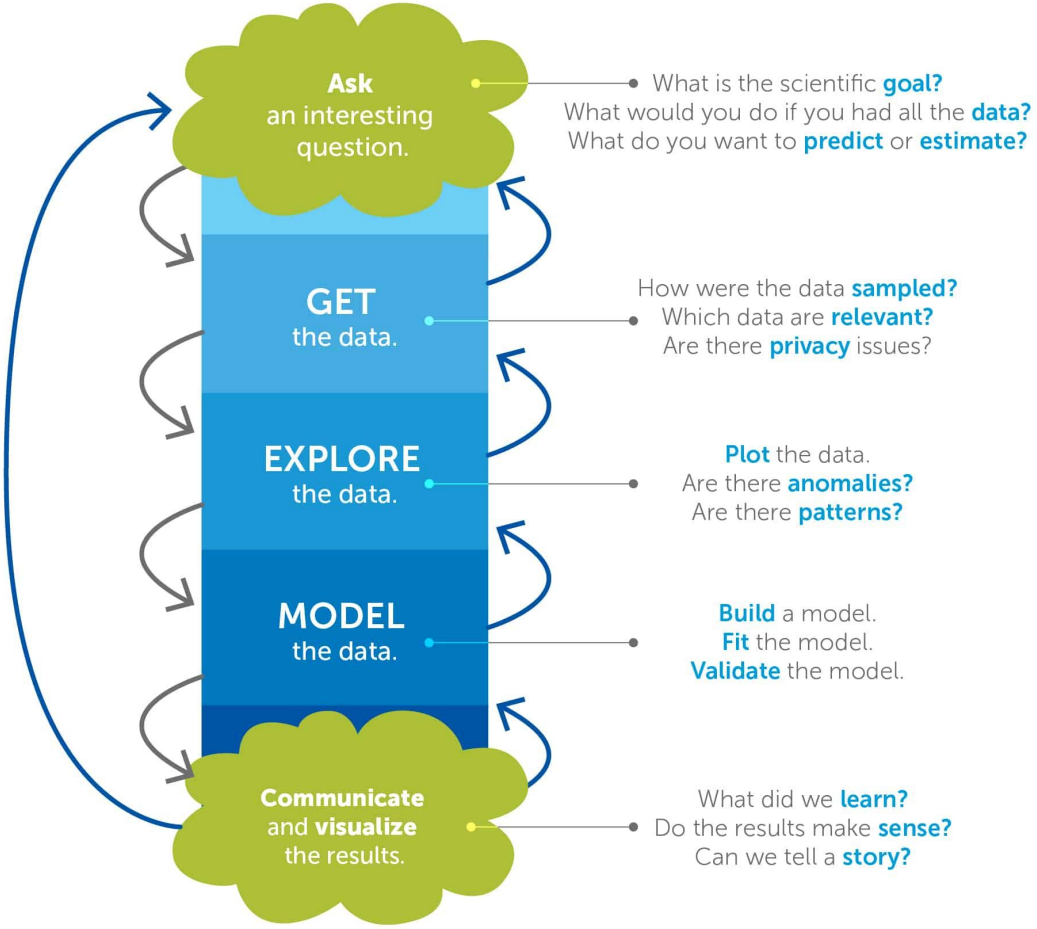


“Regardless of your exact job title, if you’re in the field of data science, you’ll be expected to be **involved in a lot of different steps** in the data-driven product development cycle. You should be ready to discover new areas to **optimize**, figure out the **metrics** that matter, find the **data** to inform these metrics, design and execute **experiments**, and **present the results** of experiments/models in concise, accurate, and convincing ways.”

The Data Science Process



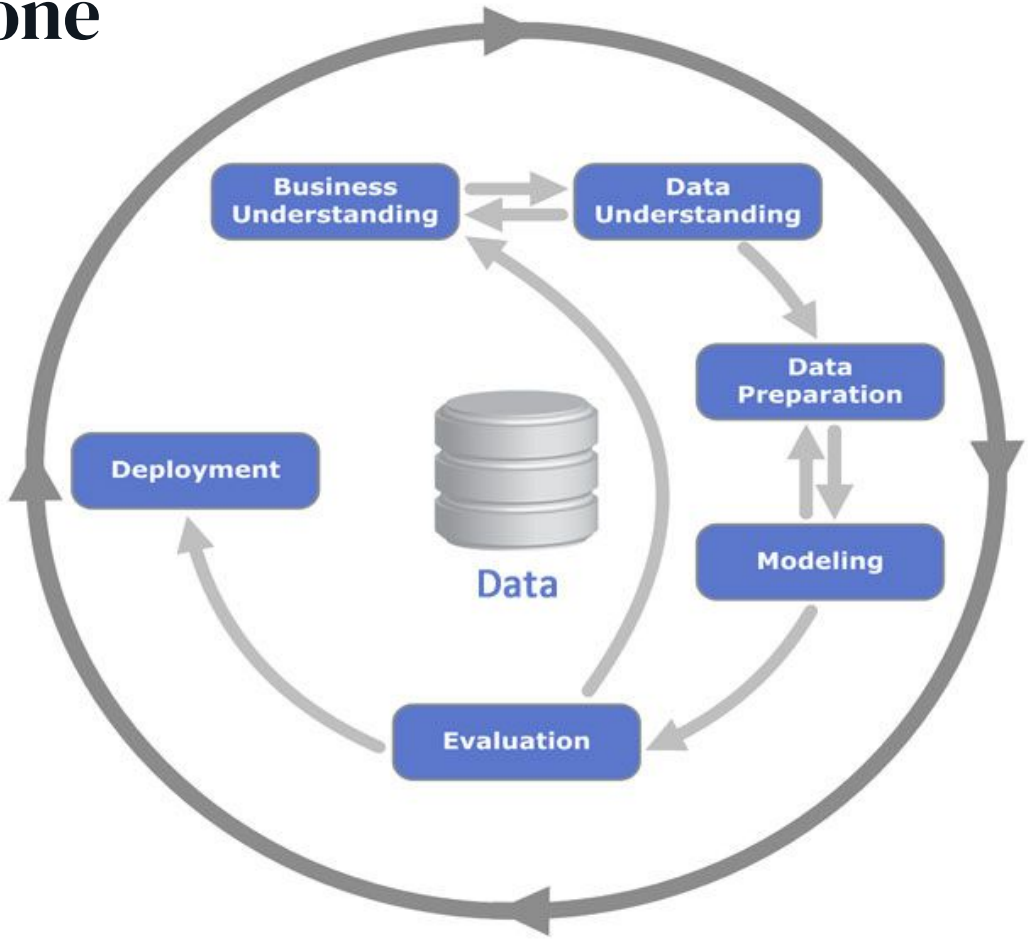
And Another



Derived from the work of Joe Blitzstein and Hanspeter Pfister,
originally created for the Harvard data science course http://cs109.org/.

You will see this one again!

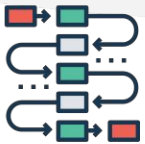
CRISP-DM Process Diagram



Source: Kenneth Jensen

Data Science Process

LEAD



OBTAIN	SCRUB	EXPLORE	MODEL	INTERPRET
O	S	E	M	N
Gather data from relevant sources	Clean data to formats that machine understands	Find significant patterns and trends using statistical methods	Construct models to predict and forecast	Put the results into good use

Originally by Hilary Mason and Chris Wiggins

The Data Science Toolkit

Data Science Toolkit - Phase 1

Languages

python™

SQL

Interfaces

jupyter

SaturnCloud

Version Control

git

GitHub

Package Control

ANACONDA.

pypi

Languages

Python

- Free, open source, versatile, powerful
- Not just for data science!
- Object-oriented (everything is an ‘object’)
- The Zen of Python

Structured Query Language (SQL)

- Connect to, change, and retrieve data from relational databases
- Developed in the 1970s, still going strong
- Many flavors

Interfaces

Jupyter Notebooks

- Streamlined document-centric interface for running and sharing code

Saturn Cloud

- Host canvas jupyter notebooks online in virtual environment

Code-Focused Text Editor

- Write text files in a code-native format
- **VS Code** is one of many that would work

Version Control

Git

- Distributed version tracking on any files
- Folder → “Repository”

GitHub

- Hosts Git repositories
- Collaborate and share code with others
- Backbone of the open source community
- Your Data Science portfolio!

Package Control

Anaconda

- Package management and deployment
- Designed with Data Science in mind
- Create and share environments

Python Package Index (PyPi)

- Database of public Python libraries
- Package installer (pip)
- Not everything is on Anaconda

Now:
Time to put this all to use!